Open Data Prioritization Toolkit

June 2015
Federal CIO Council

Table of Contents

Table of Figures

1.0 Open Data Prioritization Toolkit

1.1 Introduction

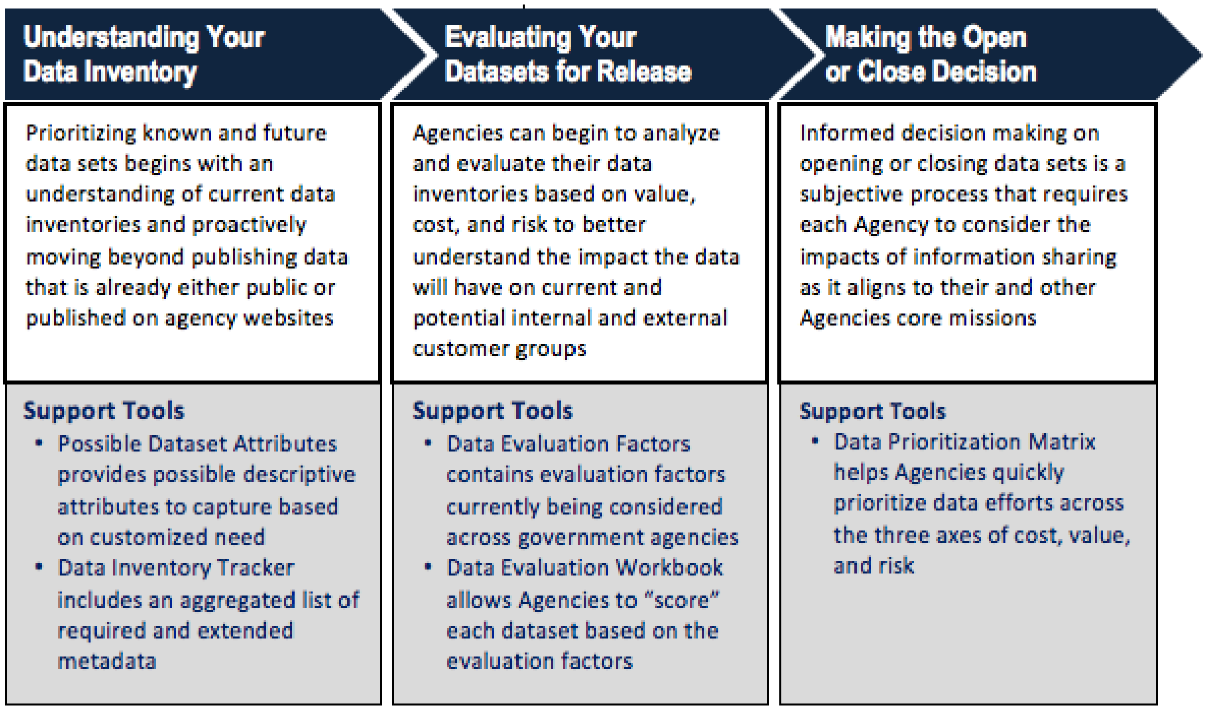
In May 2013 the Office of Management and Budget (OMB) published Memorandum M-13-13, Open Data Policy – Managing Information as an Asset, which outlined the government's vision for "making information resources accessible, discoverable, and usable by the public." In the past two years, Agencies have made significant strides in improving the amount and quality of Federal information available, culminating in the release of over 100,000 datasets to the American public.

One of the objectives of the Open Data Cross Agency Priority (CAP) goals is to "prioritize and release valuable data through public engagement." Agencies have expressed an interest in receiving guidance that would assist with prioritizing datasets to be opened. To help Agencies continue to improve the government's data sharing process, this Open Data Prioritization Toolkit was developed to assist with evaluating and prioritizing unreleased and released Agency datasets. These tools enable Agencies to evaluate what data is most critical to share with external audiences (i.e. the public and/or other Agencies); and provide a structure for gathering feedback and balancing considerations such as impacts on individuals' privacy or on national security. Visit Project Open Data to find additional information on definitions, guidance on implementation, tools, data standards and requirements, and other resources.

This toolkit is not intended to be exhaustive but rather is designed to provide Agencies with guidance and suggestions for evaluating the costs, risks, and benefits of releasing data openly; ultimately, enabling Agencies to decide how and with whom to target their data sharing initiatives among external stakeholders and between agencies. It assumes lifecycle management and documentation as a key element in increasing the overall understanding of and improving the maturity and quality of their data assets.

1.2 The Toolkit

The Open Data Prioritization Toolkit (*Figure 1*) provides a framework for identifying, evaluating, and prioritizing datasets for release. This toolkit is composed of associated tools, workbooks, and resources to provide an initial start to completion and avoid rework. It is part of the Federal Government's strategy to increase operational efficiencies, reduce costs, improve services, support mission needs, safeguard personal information, and increase public access to valuable government information through promoting interoperability and openness of Federal data. This toolkit operates at the Enterprise Data Inventory (EDI) layer and is intended to assist in comparing across one's data inventory; however it is not intended to replace any Federal Open Data reporting requirements.



Understanding Your Data Inventory

Prioritizing known and future data sets begins with an understanding of current data inventories and proactively moving beyond publishing data that is already either public or published on agency websites

Support Tools

- Possible Dataset Attributes provides possible descriptive attributes to capture based on customized need
- Data Inventory Tracker includes an aggregated list of required and extended metadata

Evaluating Your Datasets for Release

Agencies can begin to analyze and evaluate their data inventories based on value, cost, and risk to better understand the impact the data will have on current and potential internal and external customer groups

Support Tools

- Data Evaluation Factors contains evaluation factors currently being considered across government agencies
- Data Evaluation Workbook allows Agencies to "score" each dataset based on the evaluation factors

Making the Open or Close Decision

Informed decision making on opening or closing data sets is a subjective process that requires each Agency to consider the impacts of information sharing as it aligns to their and other Agencies core missions

Support Tools

- Data Prioritization Matrix helps Agencies quickly prioritize data efforts across the three axes of cost, value, and risk

Figure 1: Overview of Open Data Prioritization Toolkit

1.3 Putting the Open Data Prioritization Toolkit into Practice

Shown in *Figure 1*, the Open Data Prioritization Toolkit is categorized into phases of understanding, evaluating, and opening datasets. It also shows alignment to support tools designed to help your agency throughout each phase. The provided tools are sample templates to guide agencies as they examine datasets and customize them based on their unique requirements. The Toolkit's phases are further discussed in subsequent sections in this document with additional guidance regarding putting the Toolkit into practice.

The Open Data Prioritization Toolkit assists agencies through the process of prioritizing their data for release to the public. Depending upon the maturity of your agency's open data program, the entire toolkit may not be needed. For example, an agency with a complete record of its data inventory and 90% of its entire data inventory open to the public may pass the "Understanding Your Data Inventory" and "Evaluating Your Datasets for Release" phases, moving directly to "Making the Open or Close Decision."

2.0 Understanding Your Data Inventory

A key objective of the Open Data CAP goal is to "develop and maintain an Enterprise Data Inventory (EDI)." The EDI is part of a set of data management practices designed to help gain a clear and comprehensive view into the vast array of data assets managed by each Agency. This view, along with other tools and processes[1], will help Agencies determine and prioritize which data assets should be released. Furthermore, to fully evaluate the data assets being considered for release and the implications of releasing to the public, Agencies should consider enhancing their EDI by capturing descriptive attributes and/or metadata (e.g., Source, Format, Ownership, etc.) for the data assets as a means for dataset evaluation[2]. These additional metadata will allow agencies to more effectively evaluate their full EDI and prioritize their datasets for release.

Source

1) What is the source of the dataset?
2) Does the data source span multiple organizations
3) What, if any, processes are in place to maintain the dataset?
4) What is the associated system or program used to generate or manage the dataset?

Format and Content

5) In what format or schema does the data currently exist (e.g. XML, HTML, JSON, TXT, CSV)?
6) Is the data in a machine readable format?
7) Are the metadata[3] or attributes associated with the dataset that should be added beyond those required by the CAP goal?
8) Does a data dictionary exist for the dataset data elements?
9) Does the data include personally identifiable information (PII)[4] or other protected information that should be risk evaluated prior to release (see section 3.3.)?

Current Users

10) Who are the current internal and/or external users of the data?
11) Is the data being shared with targeted consumers or openly available?
12) Is there an understanding of how users consume and/or utilize data?
13) How many people/organizations have access to the data?
14) What is the frequency of user data consumption?
15) Has this data been requested by external users or other government agencies in the past or on recurring basis?

Frequency and Distribution

16) How often is the data collected and processed?
17) What is the lifespan of data usability?
18) What is the refresh rate of the dataset?
19) How often does the data need to be refreshed?
20) How often is the data released?
21) What mechanisms are used to distribute data?[5]

Operation and Maintenance

22) Who is responsible for managing and maintaining the data (i.e. data steward)?
23) Who is the primary data owner?
24) Is the data steward and the data owner the same person?
25) What governance processes are in place to manage the data?
26) Is there an understanding of the data lifecycle cost?
27) Are there specific dataset security requirements?

Data Integrity

28) Are there any potential issues with data credibility?
29) Is there uncertainty about data accuracy and validity?
30) Are quality control processes established and followed?
31) Are metrics calculated around data collection processes?
32) How often is data verified and validated?
33) Could a mosaic effect resulting from the release of the data reveal private or sensitive information?

Support Tools

The list of **Possible Dataset Attributes** and **Data Inventory Tracker** serves as a guide for establishing a comprehensive list of an organization's datasets. The Data Inventory Tracker reflects the Agency's Enterprise Data Inventory (EDI). Consolidating a view on one's datasets and associated metadata into a singular view is critical in evaluating datasets for releasing to the public. For Agencies with automated dataset catalogs, the concepts within the Prioritization Toolkit may be integrated into your current workflows.

3.0 Evaluating Your Datasets for Release

After establishing an accurate understanding of the data, Agencies can begin to evaluate their inventories based on value, cost, and risk to understand the impact releasing data will have on them, data consumers, and society.

3.1 Value

Prioritizing and releasing valuable data through public engagement is a priority and CAP goal. Identifying and engaging with key data consumers to help estimate the value of the multitude of federal datasets can help agencies prioritize those of highest value for quickest release, where appropriate.

All Federal agencies are required to solicit public input and reflect on how to incorporate consumer feedback into their data management practices. Agencies may develop criteria at their discretion for prioritizing the release of data assets, accounting for a range of factors, such as the quantity and substance of user demand, internal management priorities, and agency mission or strategic relevance. As consumer feedback mechanisms and internal prioritization criteria will likely evolve over time and vary across agencies, agencies should share successful innovations in incorporating consumer feedback through interagency working groups and Project Open Data.

To determine the importance of a dataset, Agencies should engage the public and gather feedback to identify stakeholders and value drivers of the data. This will allow agencies to estimate the potential impact that the data will have on its customer groups and society at-large. Value from public access to data may come in unanticipated and unexpected ways once the data is released. The following is a list of foundational questions that agencies may consider when determining the value of releasing a dataset:

Stakeholders

34) Who are the current or future internal and/or external users?
35) Who are the external government agencies?
36) What is the estimated number of users?
37) Is the data of short- or long-term value to stakeholders?
38) How frequently might the dataset be consumed?
39) How much value is derived from each data interaction?
40) Are there any limitations on data analysis and use?

Value Drivers

41) Is the data currently used to enable the performance of Agency functions or support

the Agency's mission?
42) Is the data leveraged in decision making within or outside of the Agency?
43) Does the data increase internal government efficiency?
44) Does the data improve the effectiveness of government programs?
45) What is the potential of the data to fuel innovation (e.g., enable the development of new tools)?
46) If used by secondary users, what is the potential of the data to lower costs?
47) What is the potential of the data to create economic value or growth?
48) What is the potential of the data to open up new business opportunities?
49) What is the potential of the data to catalyze new collaboration efforts?

3.2 Cost

Through considering the following questions, Agencies can begin to assess the cost of preparing data for release and maintaining the data once made public. Agencies should consider all types of monetary costs that may be incurred (e.g., amount of money required to pay for technology tools, contractor support, new positions, marketing, etc.). In addition to costs impacting operational budgets, the number of employees and the amount of time each spends dedicated to open data tasks should also be considered, as well as the tasks not being completed because resources have been reallocated to open data efforts. Any resource required for preparation and maintenance should be included in order to have a comprehensive understanding of cost.

Format

50) Will the format of the data need to be converted in order to share or use the dataset?
51) Are there definitions for the data within the dataset that ensure understanding of the data?
52) What is the estimated overall cost for data preparation?
53) What is the estimated time to prepare the data for release?

Frequency

54) How will changes be identified after the initial publication?
55) How frequently will the data require a refresh?
56) What is the estimated overall cost for data maintenance?

Review

57) Are there required processes in order to share the data?
58) How significant is the involvement of the legal department?
59) Are there regulatory (e.g. privacy, security, accessibility, etc.) concerns associated with sharing the data?

Operations and Maintenance

60) What organizations will commit human and financial resources to sharing the data?
61) What are the additional lifecycle costs for data sharing?
62) What additional technology resources will be needed?
63) What system changes need to be implemented in order to share the data? What is the estimated cost?
64) Will sharing the data require additional hosting capacity or a different hosting technology?Would de-identifying the data eliminate its utility?
65) Is a process in place for collecting public feedback on the data and what is the associated cost of maintaining that process?

3.3 Risk

Agencies need to be aware of the risk and unintended consequences associated with sharing their data. When assessing the risks associated with sharing data, Agencies should consider existing policies such as the Privacy Act of 1974[6], the E-Government Act of 2002[7],, the Federal Information Security Management Act of 2002 (FISMA)[8], Controlled Unclassified Information (CUI)[9] and Confidential Information Protection and Statistical Efficiency Act (CIPSEA)[10]. Additionally, per guidance outlined in M-13-13, Agencies are required to incorporate the National Institute of Standards and Technology (NIST) Federal Information Processing Standards (FIPS) Publication 199, "Standards for Security Categorization of Federal Information and Information Systems."[11] In order to ensure compliance with these policies and to minimize privacy or security risks associated with releasing the data, Agencies should refer to internal policies and points of contact (e.g. Privacy Officer) to help assess the risk of data sets before release. Additional questions to help guide this process are included below which are intended to assist agencies in thinking about risk generally. These questions may be addressed specifically or generalized through one or more aggregate proxy measures if not practical to answer.

Privacy and Unintended Consequences

66) Will the release of the data have any unintended consequences (e.g. discrimination against an individual / group, release of protected health information[12], or the mosaic effect)?

67) Does the data pose a security risk when combined with currently available information

Security

Note: If you anticipate that releasing data to external organizations will expose national security information, then you will most likely be unable to release the dataset.

68) Does the data disclose information regarding the security of government information or communications systems?
69) Does the data disclose information regarding physical security of government facilities (owned or leased)?
70) Does the data disclose detailed critical infrastructure information?

Other Considerations

71) Is the source of the data credible?
72) What are the potential consequences from the data being misinterpreted?
73) Are their international, foreign, or other restrictions limiting the release of data?

3.4 Support Tools

The **Data Evaluation Factors** consist of a series of questions to evaluate the datasets for release to the public. After establishing an Agency's EDI, that metadata can be transposed into the Data Inventory Tracker. As part of the Data Prioritization Toolkit, the Data Inventory Tracker can help organizations begin to evaluate their inventory based on value, cost, and risk to estimate the impact releasing the data will have on the Agency, data consumers, and society. There is no specific order in which Agencies should address these factors. At some, it may be beneficial to view and address the risk factors first as the value and cost of considering datasets may vary largely by availability of dataset identifiers.

In order to prioritize datasets for public consumption, the datasets must be assessed for their value to those outside of the organization, cost to prepare, release, and maintain, and risk to the Agency, government and citizenry if released to the public. The **Data Evaluation Workbook** provides a framework for weighting the impact and importance of particular factors to the Agency and rating each dataset by value, cost and risk, so that datasets may be compared to others. Agencies should modify any associated weighting factors based on their specific data needs and stakeholders. The open and close scoring thresholds should be set to release as much data as possible without compromising risk factors and staying within budget.

4.0 Making the Open or Close Decision

Ultimately, there is no precise formula for calculating whether or not to share information and each Agency must rely on the core values of their mission when deciding which data sets are most critical to share with the public. The Data Prioritization Matrix can assist Agencies with understanding and assessing the costs and benefits of data sharing in order to help them arrive at the best outcome for their Agency.

Support Tools

The **Data Prioritization Matrix** (*Figure 2*) is an illustration of how value, risk and cost intersect based on responses provided in the Data Evaluation Workbook (Y axis is Value, X axis is Risk, and circle size and color represent Cost). This visual representation of the datasets allows Agencies to quickly visualize their dataset prioritization for public consumption.

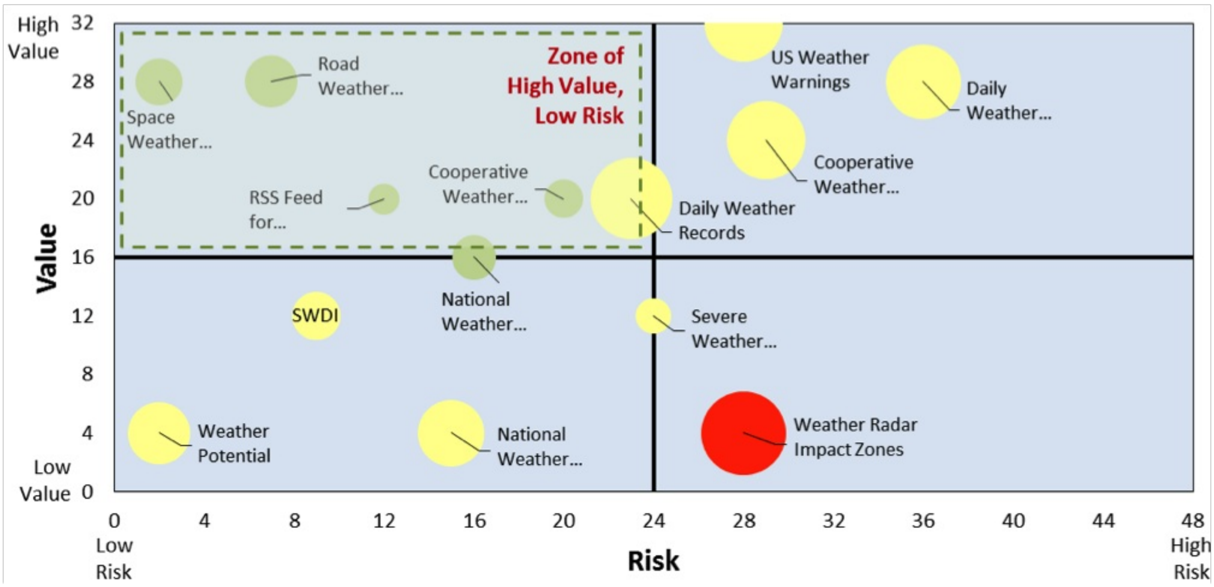


Figure 2: Data Prioritization Matrix

There are three sample scenarios illustrated:

Green: The upper left quadrant, "Zone of High Value, Low Risk," represent those datasets that provide a positive impact on consumers and the risks of releasing the dataset is minimal. The specific cost of releasing each specific dataset (size of the bubble) needs to be considered prior to releasing.

Yellow: The dataset has one or more of the following positive attributes: moderate to high value, low to moderate risk, and low to moderate cost. Therefore, the dataset should be given a lower priority for releasing.

Red: If the dataset has a moderately low value with considerable cost and risk of releasing, the dataset may be given a low priority for release. However, Agencies should rely on their organization's mission requirements and programmatic priorities to help guide data release decision making. Deciding what factors (i.e. value, cost, or risk) carry the most weight is a subjective process and will require each Agency to consider the impacts of information sharing as it aligns to the Agency's core mission.

Appendix A

Relevant Open Data Resources, Policies, Guidelines and Goal Statements

- Project-open-data.cio.gov
- Exec. Order No. 13642, 3 C.F.R. 3 (May 9, 2013) *Making Open and Machine Readable the New Default for Government Information*
- Burwell, S., VanRoekel, S., Park, T., & Mancini, D. (May 9, 2013). Memorandum for the Heads of Executive Departments and Agencies. Open Data Policy-Managing Information as an Asset. (M-13-13)
- Cross Agency Priority Goal, Open Data
- National Archives and Records Administration - Executive Order 13556, Confidential but Unclassified Information (CUI)
- U.S. Government Publishing Office (GPO): *Federal Agency Responsibilities: 44 USC 3506(d) (Information Dissemination)*
- OMB Circular A-130, Section 8.e
- Guidance for Providing and Using Administrative Data for Statistical Purposes
- Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)

Appendix B

About the Federal CIO Council Innovation Committee

The Federal CIO Council created the Innovation Committee to enable Agency mission delivery, improve customer service, maximize return-on-investment, and support emerging IT needs. The Innovation Committee focuses on relevant topics such as the use of modern technologies to deliver digital services to citizens and businesses, deployment of mobile technology within Government, modular IT development strategies, and using Federal data as a strategic resource to enable Agency mission delivery and to grow the economy

1. Federal CIO Council Case Study: The Data Disclosure Decision (Department of Education) ↩
2. Reference OMB Circular A-119 in the development, maintenance, and use of standards and specifications ↩
3. Visit project-open-data.cio.gov for metadata guidance and other references ↩
4. According to OMB M-07-1616, PII is defined as refers to information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual. The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessm33ent of the specific risk that an individual can be identified. ↩
5. Federal CIO Council Case Study: The Data Disclosure Decision (Department of Education) ↩
6. http://www.gpo.gov/fdsys/pkg/CRPT-106hrpt50/pdf/CRPT-106hrpt50.pdf ↩
7. http://www.gpo.gov/fdsys/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf ↩
8. http://csrc.nist.gov/groups/SMA/fisma/overview.html ↩
9. http://www.gpo.gov/fdsys/pkg/FR-2010-11-09/pdf/2010-28360.pdf ↩
10. http://www.eia.gov/cipsea/cipsea.pdf ↩
11. http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf ↩
12. PHI, which is individually identifiable health information held by HIPAA covered entities or their business associates, may only be disclosed as permitted by the HIPAA Privacy Rule, 45 C.F.R. Parts 160 and 164. ↩